Journal of Urban Health: Bulletin of the New York Academy of Medicine Vol. 80, No. 2, Supplement 1 2003

A Systems Overview of the Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE II)

Joseph Lombardo, Howard Burkom, Eugene Elbert, Steven Magruder, Sheryl Happel Lewis, Wayne Loschen, James Sari, Carol Sniegoski, Richard Wojcik, and Julie Pavlin

ABSTRACT *The Electronic Surveillance System for the Early Notification of Community-Based Epidemics, or ESSENCE II, uses syndromic and nontraditional health information to provide very early warning of abnormal health conditions in the National Capital Area (NCA). ESSENCE II is being developed for the Department of Defense Global Emerging Infections System and is the only known system to combine both military and civilian health care information for daily outbreak surveillance. The National Capital Area has a complicated, multijurisdictional structure that makes data sharing and integrated regional surveillance challenging. However, the strong military presence in all jurisdictions facilitates the collection of health care information across the region. ESSENCE II integrates clinical and nonclinical human behavior indicators as a means of identifying the abnormality as close to the time of onset of symptoms as possible. Clinical data sets include emergency room syndromes, private practice billing codes grouped into syndromes, and veterinary syndromes. Nonclinical data include absenteeism, nurse hotline calls, prescription medications, and over-the-counter self-medications. Correctly using information marked by varying degrees of uncertainty is one of the more challenging aspects of this program. The data (without personal identifiers) are captured in an electronic format, encrypted, archived, and processed at a secure facility. Aggregated information is then provided to users on secure Web sites. When completed, the system will provide automated capture, archiving, processing, and notification of abnormalities to epidemiologists and analysts. Outbreak detection methods currently include temporal and spatial variations of odds ratios, autoregressive modeling, cumulative summation, matched filter, and scan statistics. Integration of nonuniform data is needed to increase sensitivity and thus enable the earliest notification possible. The performance of various detection techniques was compared using results obtained from the ESSENCE II system.*

KEYWORDS *Evaluation, Nontraditional, Surveillance, Syndromes, Test bed.*

Mr. Lombardo, Ms. Lewis, Mr. Loschen, Ms. Sniegoski, Mr. Wojcik, and Drs. Burkom, Magruder, and Sari are with The Johns Hopkins University Applied Physics Laboratory; Mr. Elbert and Dr. Pavlin are with Walter Reed Army Institute of Research, Department of Defense Global Emerging Infections System.

Correspondence: Joseph S. Lombardo, The Johns Hopkins University Applied Physics Laboratory, 11100 Johns Hopkins Road, Laurel, MD 20723-6099. (E-mail: Joseph.Lombardo@jhuapl.edu)

INTRODUCTION

The enhanced version of the Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE II) is a prototype disease surveillance test bed that combines nontraditional health status indicators with new techniques to identify abnormal health conditions in the population. ESSENCE II is being developed by The Johns Hopkins University Applied Physics Laboratory (JHU/APL) under the sponsorship of the Defense Advanced Research Projects Agency (DARPA) for use in the Department of Defense Global Emerging Infections System (DoD-GEIS).

ESSENCE I is currently performing surveillance at all US military treatment facilities.[1] ESSENCE II is being developed for deployment in the National Capital Area (NCA). The ESSENCE II test bed is an evolving prototype system that is being developed while it actually performs surveillance. This approach was taken to identify needed improvements that become evident only through operation of the system. As a test bed, ESSENCE II enables the implementation and evaluation of novel surveillance concepts for its daily surveillance needs and for communicating to local health departments. The major thrust of the project is the early identification of a covert release of a deadly pathogen on an unsuspecting population, but the technology is also useful for providing an early warning of abnormal health conditions due to naturally occurring infectious diseases.

The NCA is a complex multijurisdictional region with a large military population. A significant portion of the population works and resides in different public health jurisdictions. Because pathogens do not honor jurisdictional boundaries, the sensitivity needed for the early identification of bioterrorism can be achieved only by the integration of information across the region. The ESSENCE II system is integrating both military and civilian disease indicators in a syndromic surveillance format and making the system outputs available to all public health jurisdictions in the region.

SYSTEM DESCRIPTION

ESSENCE II comprises several modular components being developed in parallel. The system architecture permits individual components to be upgraded in a nonproprietary environment with inexpensive, off-the-shelf applications software.* Figure 1 presents a functional view of the system components.

The data and the telecommunications infrastructure for acquiring and transferring the information are external to the system. The ESSENCE II modules implement the following functions:

- Policies to ensure the privacy of personal health care information
- Policies to govern the exchange of information among other surveillance or reporting systems
- A data archive
- Processes for detection of and issuing alerts about abnormalities in the indicator data
- Processes for notification of users of special events or environmental conditions that warrant changes in detection parameters

*Resources needed to operate ESSENCE II are a function of the size of the jurisdiction, the number of data sources incorporated, the threshold alerting levels the users select, and the degree of follow-up required by the user health department.

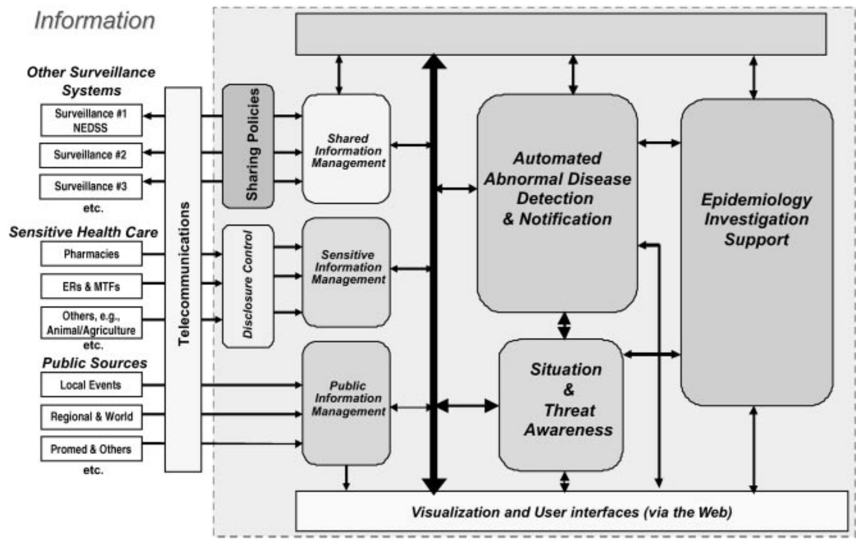

FIGURE 1. ESSENCE II functional components.

- Processes that allow the user to exploit the archive fully to identify false positives or obtain information about current or historical trends in the indicator data
- Visualization and user interfaces
- Processes for injecting simulated data for training and measuring the performance of ESSENCE II detectors and indicators

The system requires three distinct types of data. The first consists of basic information sources implicitly containing the levels of disease activity in the population. These include chief complaint data from hospital emergency rooms; *International Classification of Disease, 9th Revision, Clinical Modification* (*ICD-9-CM*) codes used for billing patient visits for private practice groups; over-the-counter (OTC) sales of pharmaceuticals that can be used for self-medication; nurse hotline calls; school absenteeism; veterinary reports (see Ref. 2 for an example of the utility of veterinary surveillance); and so on. We grouped these data streams into the sensitive health care category because they may be acquired and used only in conformance with privacy laws, corporate policies, memoranda of agreements, and the like.

These data sources also have varying degrees of specificity associated with their use in surveillance. For example, laboratory tests are more specific indicators of public health issues than are sales of OTC medications. The increase in sales may be a result of sales promotions, onset of the winter season, or stocking up in anticipation of need. Unless the uncertainties in the nontraditional data sources can be sufficiently resolved, their use must be given lower weighting in the anomaly detection process.

Providers of data in the sensitive health care category must maintain the privacy of individuals. Therefore, before ESSENCE II can use sensitive data, identifiers must be removed. If the individual records cannot be made totally anonymous, their use must be limited, and other surveillance initiatives may not be able to share them. The anonymization process can be performed before the data are sent to the

ESSENCE II system.[3] Alternatively, the sanitizing process can be performed by another surveillance system if it is willing to take on the liabilities of the provider.

Supporting information is needed to effectively use data with low specificity. These items fall into the second category of information needed to understand and operate the ESSENCE II system—publicly available information. For example, information about local endemic disease, sales promotions, and even weather events is important to support the use of OTC medications as an early indicator of the presence of disease. This information in most cases is available via electronic media. Likewise, the occurrence of high-profile events in the community may change detection and alerting thresholds.

The third category of information consists of the products of external surveillance activities in the NCA. Many of these activities generate results that could be useful for increasing the sensitivity, specificity, and timeliness of alerts from the ESSENCE II system. Agencies conducting the surveillance activities must agree on which data elements can be shared. Integral to the ESSENCE II system will be a set of rules that implement the policies agreed to by the providers and recipients of sensitive data.

SENSITIVE HEALTH INDICATOR DATA SOURCES

For ESSENCE II to achieve the desired early warning performance,* it must evaluate all reasonable data sources. Therefore, a primary objective of the ESSENCE II project is to identify sources of data that contain early indicators of abnormal disease in the population. Many of the data sources are unconventional when compared with past or present surveillance activities undertaken by public health entities. Data sources can be segmented into three categories: traditional, nontraditional clinical, and nontraditional nonclinical. The traditional gold standard is the confirmed laboratory result, but this data source may not provide the timeliness needed to respond to a widespread outbreak caused by a covert attack with a disease delivered as a weapon.

Nontraditional clinical data are obtained from encounters with health care professionals. These data are potentially crucial to early detection because the early presentation of diseases caused by a biowarfare attack is likely to resemble common illnesses such as influenza.[4] Syndromic surveillance, using the ESSENCE syndrome groups when possible, is applied to these data. Data sources in the nontraditional clinical category include 911 calls, nurse hotline calls, poison center calls, visits to private practice physicians and military clinics, requests for laboratory work, emergency room visits, and prescription medications.

Daily counts are placed into the following syndrome groups:

- Respiratory
- Gastrointestinal
- Fever
- Dermatological, hemorrhagic
- Dermatological, infectious
- Neurological
- Coma

*One of the ongoing research areas in ESSENCE II is the determination of the ideal alarm levels. At present, alarms are investigated within the project and forwarded to the health department when they cannot be resolved.

Each of these groups is defined by a specific set of *ICD-9* codes. In conducting surveillance of symptoms that occur in the early stages of disease, the system monitors the occurrence of common diseases. The daily syndrome group counts usually reflect normal background levels. The key to identifying abnormalities in syndrome group levels effectively is the ability to model and estimate normal background levels, which vary as a function of season and the normal evolution of endemic disease strains.

All demographic information that could be used to identify individuals is being removed from data files before they are sent to the ESSENCE II test bed. One problem with deidentified nontraditional clinical data is that a single case of illness could show up in several of the data streams used for surveillance. For example, a call to a nurse hotline could result in a visit to a private practice physician, a request for a laboratory test, and a medication prescription. Without a way to link the data sources by individuals, these dependencies tend to degrade the performance of alerting algorithms. Military patient records, as well as those of many health maintenance organizations, can be used to link data sources because they contain identifiers. The other approach is to find data sources that are totally independent. Capturing surveillance data on diseases that are present in both animal and human populations provides an independent data source that can be used to reduce false positives. These independent sources may not always be available.

Nontraditional, nonclinical data contain disease indicators that are not reported as the result of an encounter with a health care professional. Included in this category are absenteeism and the purchase of OTC medications. Such data cannot easily be grouped by syndrome. OTC medications can be grouped as antiflu, antidiarrheal, or the like, but absentee records do not typically indicate the causes of absence. It is also difficult to determine from the data elements available how many independent occurrences of illness exist in these sources. Despite this drawback, Magruder[5] and Sari[6] have shown that, on average, increases in OTC sales of antiinfluenza medications have preceded increased activity in emergency rooms by up to 4 days.

Figure 2 from Sari[6] compares the sales of flu medications with emergency room activity for fever during the 2000–2001 winter season. The curves represent the activity levels for a major drugstore chain compared with all of the emergency rooms in the city of Baltimore, Maryland. The curves have been normalized by dividing the daily levels by their seasonal mean. The figure is typical of other comparisons showing that the OTC sales of flu medications lead emergency room activity by several weeks. It is difficult to separate actual illness represented by the sales versus stocking up for the upcoming cold and flu season. Several years of data are needed to understand the sales patterns of these medications.

In the course of developing ESSENCE II, data sources are continually being added and evaluated. The goal is to obtain a sample density that will support early recognition of abnormal patterns. Currently, 100% of the clinical visits of military and their dependents are included.

ELECTRONIC DATA COLLECTION AND FORMATTING

Disease surveillance activities requiring daily feeds from several data sources must rely on modern information technology and telecommunications. Automated collection, transfer, formatting, processing, and visualization are needed to ensure continuous operations. Further, the processes implemented must fit into the business rules and privacy policies of the organizations supplying the data.

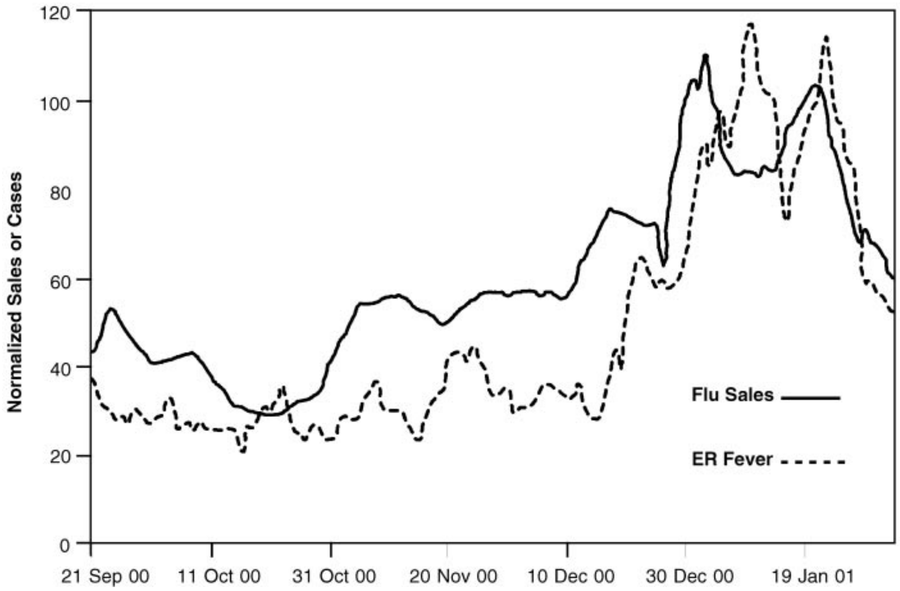

FIGURE 2. Comparison of over-the-counter antiinfluenza medication sales in Baltimore with hospital emergency room activity for fever in 2000–2001 flu season.[6] (Used with permission.)

Wojcik[7] and Sniegoski[8] have defined methods for electronic acquisition and formatting of sensitive health indicator data. We selected the automation hospital emergency room data as an example. Chief complaint text fields are typically captured and archived on an emergency room log. Scheduled queries can be made to this archive to generate an electronic report that includes a limited set of data fields, which satisfies privacy policies while capturing those fields needed to perform surveillance. The electronic report is encrypted, sent to a secure file transfer protocol (FTP) site, and placed in a unique location set up for the hospital's data. The surveillance system continually polls the site to update information and transfers the encrypted record into the surveillance archive.

As soon as the new record is received in the archive, a 13-step natural language parsing process sanitizes the chief complaint text and converts it into syndrome groupings. Once converted to this common format, the information is available for use or for other surveillance activities. Within minutes of the query to the hospital emergency room electronic log, the system can forward counts of the syndrome groups to the participating hospital, state, and county surveillance activities.

DETECTION OF ABNORMAL HEALTH CONDITIONS

The ESSENCE II project has investigated several analytical processes using heterogeneous data types for detecting abnormal syndromic levels. Models are needed to estimate normal background behaviors effectively for each type. Abnormal disease patterns can be determined by constructing detectors that operate on input time series of counts.

Detecting an abnormal syndrome level at an individual military facility is best accomplished by a temporal detection process. The ESSENCE I surveillance system uses temporal detectors to alert preventive medicine officers to the presence of these abnormal levels at their facility.[9] The ESSENCE program has investigated such statistical techniques as odds ratios[10] and autoregressive modeling,[11] cumulative summation[12] techniques from industrial quality control, and matched-filter techniques[13] used in radar and sonar signal processing.

Surveillance using purely temporal techniques depends on the choice of geographic regions with counts that are included in the time series. If the aggregate region is too small or not well chosen, the counts may be too small for accurate background representation, and outbreak cases may be excluded. Conversely, if the region contributing to the time series is too large, early cases of an outbreak may be lost in the noise. In practice, the temporal algorithms are usually run at the county level. The temporal detector currently used in ESSENCE II is the project-modified exponentially weighted moving average technique, which performs well for large and small daily counts with low false alarm rates. ESSENCE II also incorporates the algorithms of the Early Aberration Reporting System (EARS) of the Centers for Disease Control and Prevention.[14] The user may select either.

Spatial processing techniques are applied to avoid the problem of spatial preselection bias. Most data available to ESSENCE II can be resolved to only the patient ZIP code. The Health Insurance Portability Accountability Act of 1996 prohibits greater resolution because it specifies that local jurisdictions may limit the use of information that can identify health care records of individuals. The primary spatial approach in ESSENCE II has been an enhanced use of the Kulldorff scan statistic,[15,16] as implemented in the SaTScan software available from the National Cancer Institute. This method combines a likelihood ratio statistic developed by Kulldorff with a cluster analysis technique that finds clusters of maximum likelihood regardless of location or extent. Several modifications have been necessary for applying the method to the various data sets of ESSENCE II; because the spatial data are not generally proportional to the populations of the underlying ZIP codes, modeling or data history is used to calculate expected ZIP code counts. Substantial data analysis has been done to reduce the false cluster rate.

Burkom and Elbert[17] applied the Kulldorff statistic to multiple data sources in ESSENCE II by treating them as covariates while using whatever spatial information is available in each source. Figure 3 presents an example of this application with three data sources: OTC sales of antiflu medications, school absentee counts, and office visits with diagnoses coded as an influenza-like illness, a subgroup of the ESSENCE respiratory syndrome. The figure represents a test case for which a significant cluster was found near Annapolis, Maryland. The spatial information used for this analysis included patient ZIP codes and exact store and school locations. An advantage of this method is that additional sources or improved spatial information are readily incorporated.

The modified scan statistic produces approximate clusters of space-time interaction. Its clusters may also be used to reduce the preselection bias in the temporal methods noted above or in more elaborate spatial-temporal alerting methods. Given the appropriate surveillance focus, a time domain matched-filter method[13] has shown the potential for excellent sensitivity as a regional anomaly detector.

INTERNET-BASED INFORMATION DISTRIBUTION

A basic function of the ESSENCE II system is to deliver alerts and surveillance information to the military and civilian public health authorities in the NCA. The system provides detector outputs as well as the details of individual data streams via secure Web sites. Figure 4 provides an example from the ESSENCE II site similar to

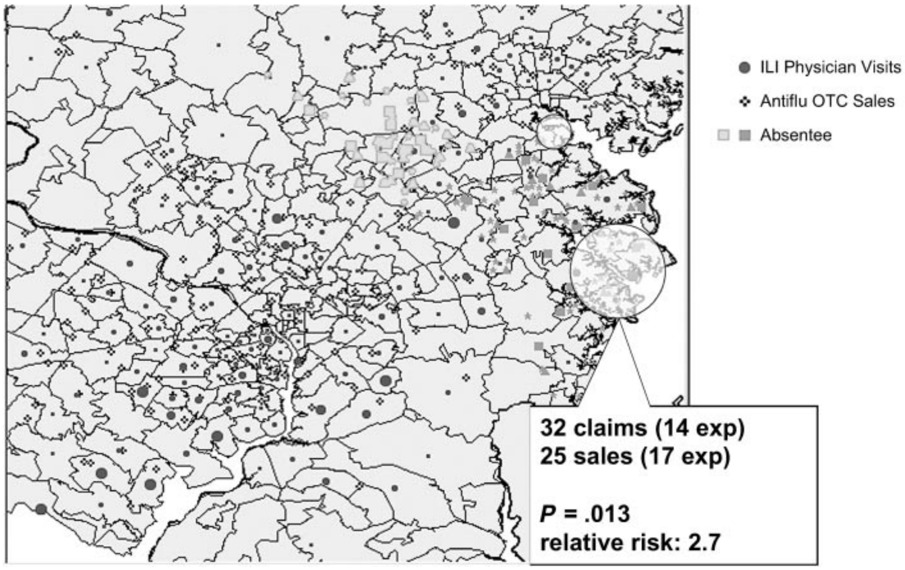

FIGURE 3. Cluster formation employing the scan statistic with three data sources.

that generated by Loschen.[18] Information is provided in many separate information layers. Layering was implemented to facilitate the distribution to the various users. Separate user names and passwords are provided so that ESSENCE II can recognize each user and provide only the data the user is authorized to view. For example, a user who logs on from an emergency room may be able to see only the emergency room data from his or her jurisdiction, whereas a user recognized as a director of epidemiology would have access to all the information within his or her jurisdiction as well as the shared information from the surrounding jurisdictions in the region.

Each ESSENCE II layer is divided into sublayers. For many of the sensitive health care data layers, the sublayers are the ESSENCE syndrome groupings. Sublayers for OTC product groupings contain antiinfluenza and antidiarrheal medica-

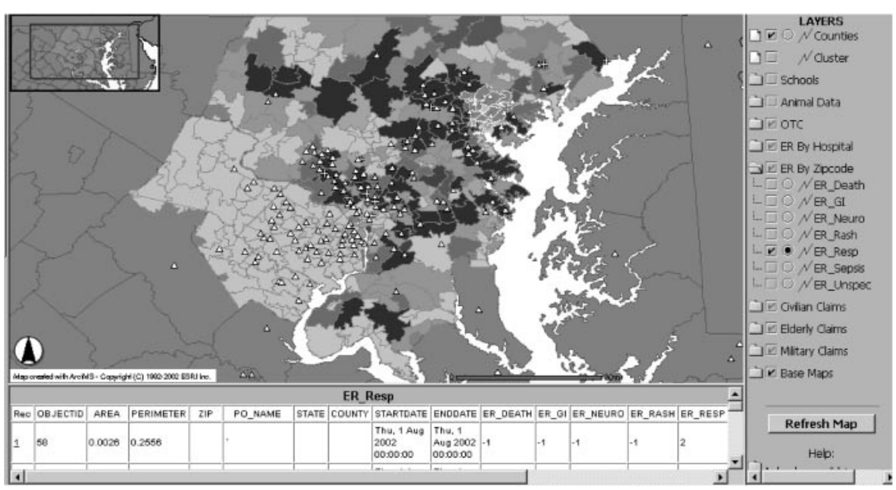

FIGURE 4. Web-based information distribution.

tions. Sublayers for school absenteeism data are the sets of elementary, middle, and high school records. The cursor can be used to select individual or groups of ZIP codes for temporal plotting of individual data elements. ESSENCE II contains lists of data indicators prioritized by the degree of the anomaly.

An authorized user can access all archived information and can select various plotting formats for presenting the data to help resolve false positives or initiate an investigation when needed.

EVALUATION OF ESSENCE II PERFORMANCE

A major difficulty in the development of a disease surveillance system is lack of experience with authentic outbreak data with which to design the system. ESSENCE II has relied on two separate approaches for evaluating its early alerting capability. The first is its ability to identify unusual endemic disease events. Large seasonal events are relatively easy to detect (allergies, influenza, etc.) compared with smaller events that are more contained in time and space.

A second method for determining performance is to synthesize a variety of different bioterrorist events. This requires an understanding of susceptibility, infectivity, symptoms at onset, human behaviors, and many more physical parameters.[19] Given the incomplete understanding of the effects of many of the possible pathogens, constructing an accurate outbreak model that includes estimates for indications in all of the ESSENCE II data streams is very difficult. The validity of the actual estimates obtained can be questioned, but estimates are needed if a solution is to be found. These estimates can be refined as additional knowledge to model the parameters properly is obtained.

The approach taken by ESSENCE II is to develop several outbreak scenarios to test the performance of the processing used in the system. The method consists of the following steps:

1. Select a time, location, pathogen, and mode of dispensing the material with a specific terrorist objective in mind.
2. Review the literature for previously published information on infectivity, incubation period, onset distribution, symptoms at onset, the change in symptoms into the acute phase of the disease, and so on. Note that ESSENCE II has relied on previous work by Sartwell[20] to estimate the distribution of the onset of disease symptoms.
3. Estimate the percentage of the infected population fitting into socioeconomic classes and age brackets.
4. Estimate the behaviors of economic classes and age brackets from analysis of behaviors during previous influenza seasons.
5. Using steps 3, 4, and 5, create a temporal and spatial model of the additional numbers of cases, products sold, persons absent, and so on.
6. Merge additional cases with actual data streams obtained previously during the same time of the year as the simulated event.

The result is a series of real data streams with a simulated outbreak superimposed. The alerting algorithms use these data streams as a test case for evaluation. The number of infected people and the units added to each data stream are treated as parameters not only to test the performance of the algorithms, but also to assess the value added of each data source. Figure 5 provides an example of detector

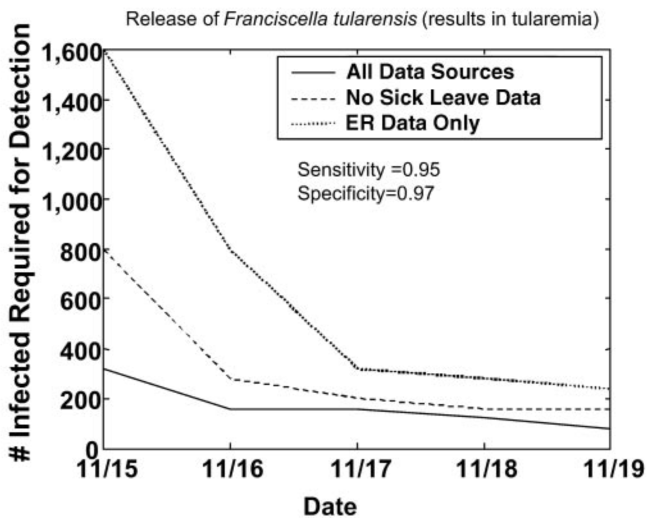

FIGURE 5. Parametric evaluation of detector performance. Indicator data: (1) sick leave; (2) over-the-counter medications; (3) physician office visits; (4) emergency room visits.

performance as a function of the number of infected people and the number of data sources used in the detection process. The number of infected people is varied to achieve a detector performance with a sensitivity of 0.95 and a specificity of 0.97.

The data set for this study included four data types: hospital emergency room respiratory syndrome counts, office visit respiratory counts, OTC influenza medication sales, and school absentee totals. Figure 5 is from a targeted surveillance scenario in which a matched-filter algorithm was applied over a fixed geographic area.

The scenario shown in the figure includes a large percentage of school-aged children in the infected population, with onset of the symptoms occurring during the week when these children would normally be at school. For this scenario, absenteeism was a major contributor to the early detection of the abnormality. It permitted the abnormality to be identified with 300 infected when used in conjunction with the other data sources. Similar performance with only emergency room syndromic surveillance would require an infected population of 1,600. Similar performance could be achieved 2 days later with emergency room data alone. The performance outcome is a direct result of the scenario chosen (time, location, incubation period, etc.). If school had not been in session when the onset occurred, this result could not have been achieved. A scenario with onset occurring on a weekend, in the inner city, or in the summer would favor the emergency room indicator. For a surveillance system to be effective, it must evaluate each indicator source in the context of a wide variety of scenarios. Additional work in this area is ongoing.

REFERENCES

1. Pavlin JA. Rapid detection of disease outbreaks. *Army AL&T*. November–December 2001;47–48.
2. Eidson M, Komar N, Sorhage F, et al. Crow deaths as a sentinel surveillance system for West Nile Virus in the northeastern United States, 1999. *Emerg Infect Dis*. 2001;7: 615–620. Available at: www.cdc.gov/ncidod/EID/vol7no4/eidson2.htm. Accessed September 1, 2002.

3. Sweeney L. Guaranteeing anonymity when sharing medical data, the datafly system. In: *Proceedings, American Medical Informatics Association*. 1997;4(suppl):51–55.
4. Franz DR, Jahrling PB, Friedlander AM, et al. Clinical recognition and management of patients exposed to biological warfare agents. *JAMA*. 1997;278:399–411.
5. Magruder SF. Evaluation of potential data sources for surveillance. Poster presentation at: National Syndromic Surveillance Conference, New York Academy of Medicine; September 23–24, 2002; New York, NY.
6. Sari JW. *Lagged Correlations Between Weather, Product Sales, Emergency Room Visits, and Claim Syndrome Groups in the Baltimore–Washington areas for 1999–2001*. Laurel, MD: The Johns Hopkins University Applied Physics Laboratory; September 9, 2002. Internal report STX-02-025.
7. Wojcik RA. *Automated Data Ingestion*. Laurel, MD: The Johns Hopkins University Applied Physics Laboratory; September 9, 2002. Internal report STJ-02-010.
8. Sniegoski CA. *Free Text Chief Complaint Text Processing*. Laurel, MD: The Johns Hopkins University Applied Physics Laboratory; August 9, 2002. Internal report STJ-02-009.
9. Elbert E, Burkom HS. Temporal alerting algorithm methodology for ESSENCE syndromic surveillance data. Poster presentation at: National Syndromic Surveillance Conference, New York Academy of Medicine; September 23–24, 2002; New York, NY.
10. Gorgis L. *Epidemiology*. 2nd ed. Philadelphia, PA: W. B. Saunders Co.; 2000.
11. Gallant AR, Goebel JJ. Nonlinear regression with autocorrelation errors. *J Am Stat Assoc*. 1976;71:961–967.
12. Tillett HE, Spencer IL. Influenza surveillance in England and Wales using routine statistics. *J Hyg Camb*. 1982;88:83–94.
13. Burkom HS, Lombardo JS, Newhall BK, et al. *Automated Alerting for Bioterrorism Using Autonomous Agents*. Laurel, MD: The Johns Hopkins University Applied Physics Laboratory; March 2001. Report STD-00-373.
14. Lawson B, Fitzhugh E, Hall S, Garcia M, Hutwagner L, Seeman GM. Implementing the CDC Early Aberration Reporting System (EARS): a front-line perspective from the Knox County (TN) Health Department. Poster presentation at: National Syndromic Surveillance Conference, New York Academy of Medicine; September 23–24, 2002; New York, NY.
15. Kulldorff M, Nagarwalla N. Spatial disease clusters: detection and inference. *Stat Med*. 1995;14:799–810.
16. Kulldorff M. A spatial scan statistic. *Commun Stat Theory Meth*. 1997;26:1481–1496.
17. Burkom HS, Elbert E. Biosurveillance applying scan statistics with multiple, disparate data sources. Poster presentation at: National Syndromic Surveillance Conference, New York Academy of Medicine; September 23–24, 2002; New York, NY.
18. Loschen WA. *User Interfaces for ESSENCE II*. Laurel, MD: The Johns Hopkins University Applied Physics Laboratory; September 9, 2002. Internal report STJ-02-008.
19. Happel Lewis SL, Cutchis PN, Babin SM. *Simulation of Pneumonic Plague in Montgomery County, Maryland*. Laurel, MD: The Johns Hopkins University Applied Physics Laboratory; September 2000. Internal report STJ-02-001.
20. Sartwell PE. The distribution of incubation periods of infectious disease. *Am J Epidemiol*. 1995;141:386–394. Originally published in *Am J Hyg*. 1950;51:310–318.